2 / PRTS [10917/11]

METHOD FOR RECOGNIZING SPEECH USING A GRAMMAR

~~Description~~

Field of the Invention

The present invention relates to a method for recognizing speech from word sequences assembled from multiple words of a given vocabulary.

Related Technology

The error rate for recognition of continuously spoken speech that permits any desired combination of all words rises considerably by comparison with individual word recognition. To counteract this, knowledge about permissible word sequences is stored in so-called language models, and used during recognition in order to reduce the number of word sequences.

Language models are usually defined as so-called N-gram models, N designating the depth of the model; in other words, N successive words within a word sequence are taken into account during the current evaluation. Because the complexity of the recognition process rapidly rises with increasing values of N, digram (N = 2) and trigram (N = 3) language models are the ones principally used.

German Patent No. 195 01 599 C1, which is hereby incorporated by reference herein, describes, in addition to various previously known methods for speech recognition, a method that allows the storage in a digram language model of phrases having fixed syntax and any desired length N. The method integrates knowledge about the syntax of permitted phrases (word sequences) into the language model, and is therefore also referred to as a "syntactic digram." An essential element for integrating syntax into the language model is the indexing of words that occur more than once in different phrase constellations. As a result, the speech recognition system is identical with and without integrated syntax.

With the severe limitation of the permissible word sequences and a limited number of permitted phrases, the speech recognition system operating according to the syntactic digram language model achieves a high recognition rate but is also usable only if syntactic limitations can be reliably defined and adhered to, for example in the case of short commands, date or time inputs, and the like. If the number of

permitted word sequences is large, however, complete definition of the syntax is very laborious; and in situations where spontaneously formulated word sequences also need to be recognized, and in which there is no guarantee that syntactic limitations will be observed, recognition using a strictly syntactic language model is of only limited suitability.

It is therefore ~~the object~~ of the present invention to describe a method for recognizing speech that offers an expanded area of application compared to existing methods, with a good recognition rate.

~~The present invention is described in Claim 1. The dependent claims contain advantageous embodiments and developments of the present invention.~~

The combined utilization of two different recognition methods, in particular having different degrees of syntactic limitation, preferably of recognition methods based on a language model with unique syntax on the one hand, and of a statistical N-gram language model on the other hand, results, surprisingly, in a considerably expanded area of application, yielding a variety of possible combinations. ~~What is essential about the combination is that~~ successive word sequence segments of a cohesive word sequence are processed using different recognition methods. Depending on the area of application, a different division of the overall word sequence into segments, and use of the various recognition methods, may be advantageous. In this context here and hereinafter, what is meant as "words" is not only words in the linguistic sense as sound sequences having a demonstrable conceptual content; "words" are rather to be understood in general as sound sequences processed integrally in the speech recognition system, for example including the speaking of individual letters, syllables, or syllable sequences without a specific conceptual assignment.

When a word sequence is divided into one or more segments, it is possible in particular to predefine at least one segment in terms of position and/or length. A predefined segment of this kind can be positioned, in particular, at the beginning of a word sequence, and can also have a fixed length in terms of the number of words that it encompasses. Advantageously, the recognition method with the integrated unique syntax can then be allocated to this segment. Because of the limited length of the segment, the outlay in terms of syntax definition and processing using the recognition

method with integrated unique syntax remains within acceptable limits. At the same time, the number of plausible word sequences can be considerably limited because the syntax is defined and is taken into account in the first segment. One advantageous field of application of this is the input of concepts by spelling. For example it is possible to recognize several tens of thousands of different city names by spelled-out speech input, with a surprisingly high recognition rate and little outlay, by combining an initial segment of fixed length that is processed on the basis of a recognition method with integrated unique syntax, and further processing of the speech input following that segment using a statistical N-gram recognition method, in particularly a digram or trigram recognition method. If exclusively a recognition method with integrated unique syntax were used, the outlay for syntax integration and process would greatly exceed tolerable limits. On the other, the exclusive use of a statistical language model in such cases would yield inadequate recognition rates.

Other advantageous examples of the segment-wise utilization of a recognition method with integrated unique syntax include word sequences with date or time information, whose word environment can then advantageously be processed with a statistical language model.

It is particularly advantageous if a statistical language model is combined with a language model with integrated syntax limitation even for the recognition of word sequences in which recurrent characteristic terms or phrases can be expected. In this context, the statistical recognition method is preferably used as the standard procedure; and if the word flow is monitored in a manner known per se for specific terms or phrases ("word spotting" or "phrase spotting"), it is possible, when such terms or phrases are detected, to initiate a segment in which speech recognition is performed using the detection method with integrated unique syntax. This segment can possess a fixed or variable length, which in particular can also be adapted to the respective term or phrase. After the completion of this segment, if the word sequence continues, it is then possible to change back to the standard recognition method with statistical word sequence evaluation.

For the recognition method with integrated unique syntax, it is preferable to use the syntactic digram recognition method known from the existing art cited initially. For the statistical speech recognition method with word sequence

evaluation, a digram recognition method is then also advantageous for application of an integral speech recognition system. On the other hand, a statistical recognition method with a higher value of N yields an improved detection rate, but also requires greater processing outlay. An advantageous compromise is to use a trigram recognition method for the statistical recognition method; a preferred embodiment of the present invention provides for performing recognition with the information volume of a trigram recognition method, in the form of digram processing.

Brief Description of the Drawings

The present invention is illustrated in even further detail below with reference to preferred exemplary embodiments referring to the drawings, in which:

Figure 1 shows a simple processing sequence diagram using the example of a spelled-out speech input;

Figure 2 shows a network graph according to the existing art;

Figure 3 shows the graph of Figure 2 with additional syntactic limitation;

Figure 4 shows the beginning of the graph of Figure 3 utilizing the present invention; and

Figure 5 shows an expanded example based on the principle of Figure 4.

Detailed Description

The example selected for explanation of the present invention with reference to the Figures drawing is spelled-out speech input of city names. The lexicon of a spelling recognition system to be used for this purpose comprises approximately 30 letters as well as a few additional words such as "double" or "dash." The list of city names contains, for example, several tens of thousands of entries, so that complete storage of the unique syntactic information (in this case the letter sequences) would increase the magnitude of the lexicon containing the syntactic information, and the computing time required for recognition, to unacceptable levels.

The sequence diagram sketched in Figure 1 for the recognition of a spelled-out entry with no parameters of any kind indicates, by way of the arrows, that proceeding from a Start node, the word sequence (which, in the particular example selected, is a sequence of individually pronounced letter names) can begin with any one of the letters provided for, and any letter can be followed by any other letter unless the word sequence has already ended, as represented by the End node.

In the conventional network graph depiction, network paths are shown, for example, for the German city names Aachen, Aalen, and Amberg. As set forth in German Patent No. 195 01 599 C1 already cited as existing art, in a network graph of this kind the identical word nodes (letters) occurring at various positions of the network yield not only the plausible word sequences provided for by the network paths, but also in a plurality of nonsense word sequences that nevertheless qualify as permissible according to the language model.

To eliminate this problem, German Patent No. 195 01 599 C1 proposes to use indexing in order to distinguish those word nodes which occur more than once in the network. Indexing makes all the word nodes of the network unique, and for each word node it is possible to indicate completely, as the syntax describing the totality of all permissible word sequences, the permissible subsequent word nodes. Especially in the case of spelled-out input of terms from a long list of terms, the ambiguity of the network graph without indexing is enormous.

Based on the example of Figure 3, Figure 4 depicts the procedure according to the present invention. What is selected, for purposes of illustration, is a variant of the present invention in which at the beginning of the word sequence, a segment of constant predefined length is processed using a recognition method with unique syntax integration, and a changeover is then made to a statistical recognition method with word sequence evaluation. The basis for the recognition method with unique syntactic limitation is a syntactic digram recognition method. The length of the introductory segment at the beginning of the word sequence is assumed to be $k = 3$ words. It is assumed for the subsequent segment of the word sequence, whose length is a priori not known or limited, that a statistical recognition method with word sequence evaluation, and with the information depth of a trigram method, will be used. In order to illustrate a particularly preferred embodiment of the present invention, a description will also be given of processing of the trigram information using a digram recognition method, by the fact that the information volume of three words (word triplet) present inside the trigram window is divided into two overlapping pseudowords (word doublet) that each comprise a combination of two successive words of the underlying trigram window.

In the example sketched in Figure 4, proceeding from the Start node, at the

beginning of a word sequence a syntactic digram recognition method is applied in a manner known from the existing art. For the city names entered in Figures 2 and 3 as network paths:

AACHEN

AALEN

AMBERG,

this means that the first three individually spoken letters

A A C

A A L

A M B

are processed with the syntactic digram recognition method. For processing of the subsequent word sequence segment using a trigram recognition method, it is advantageous if the information from the first segment can also already be evaluated as history for the beginning of the second segment. For processing with the information depth of a trigram, this means that the letter sequences

A C H E N

A L E N

M B E R G

of the information should advantageously be available with trigram information depth. The processing in the second segment of the word sequence entered in spelled-out fashion therefore advantageously also includes the last two letters of the first segment.

It is particularly advantageous if the same speech recognition system can be used in all successive segments. For this purpose, in the second segment the information present with trigram information depth is now processed using a digram recognition method. This is done by reshaping the word triplet of the trigram window, which is shifted stepwise sliding fashion along the word sequence, into a pseudoword doublet in which each two adjacent words of the word triplet of the trigram window are combined into one pseudoword. For the examples selected, the result is thus a sequence of pseudowords of the following type:

AC CH HE EN

AL LE EN

MB BE ER RG,

in which each two successive pseudowords (letter pair) contain the speech information of a word triplet from one trigram window. Reshaping the word triplets into pseudoword doublets makes possible digram processing, which takes into account only two successive pseudowords in each case, while retaining the trigram information depth. Because digram processing is used in the second segment as well, the design of the speech recognition system remains the same over the entire word sequence.

For the transition from the first segment with processing based on a syntactic digram recognition method to the second segment with processing based on the pseudoword digram recognition method without syntactic limitation, it is advantageous if, in the first segment, the last word node has added to it the information of the previous word node; this results, in the first segment, in a sequence of word nodes (letters) of the following kind:

A A AC

A A AL

A M MB;

the last word node once again constitutes a pseudoword with the information of the previous node.

Figures 5 depicts a portion, configured using this principle, of the network graph for the examples also selected in Figures 2 and 3. Proceeding from a Start node, in the first segment the network is built up with individual word nodes (individual letters) which then, at the transition to the second segment, transition into pseudoword nodes each having the information volume of two successive letters. The transitions between the pseudoword nodes are evaluated, in a manner known per se, on the basis of learning samples. The resulting network graph comprises a combination of the two different recognition methods. Despite the considerably greater number of distinguishable pseudowords as compared to the number of different letters, dispensing with continuous application of a syntactic limitation over the entire network results in a considerable reduction in processing outlay, with a high recognition rate.

In the example of Figure 5, arrows from each of the pseudoword nodes to the

End node indicate that even after only a portion of the entire word sequence, the speech input may already be sufficient for allocation of a term from the predefined list. In a recognition system, this can be implemented by the fact that once the number of terms considered relevant after input of a portion of the word sequence has been sufficiently limited, the recognition system offers a selection of terms (on a display, for example) so that input can thereby be shortened.

The present invention is not limited to the exemplary embodiments described, but rather can be modified in various ways in the context of the capabilities of one skilled in the art. In particular, the degree to which syntactic information is taken into account in the second method is variable.